Adaptive Consensus

From diagnostics to training dynamics

A 20.4M-parameter language model.
Four local models. One investigation.

01 GRADIENT NOISE

02 CURVATURE

03 TRAINING DYNAMICS

A controlled, small language-model experiment

MODEL

20.4M parameters
8 layers · width 320 · 5 heads
FFN width 896 · vocabulary 32,000

DATA & BUDGET

C4 · TinyLlama tokenizer
Context 1,024 · ≈408M tokens
40 virtual epochs · seed 42

OPTIMIZATION

AdamW
$(\beta_1, \beta_2) = (0.9, 0.999)$
Weight decay 0.1
Gradient clipping at 1.0

LEARNING-RATE SCHEDULE

5% token warmup
Cosine decay to 10% of peak

Same architecture, data source, global batch, and ≈408M-token budget.

The tied embedding / output matrix accounts for 10.24M parameters.

Four local models, independent optimizer states

SYNCHRONOUS

32,768-token batch

↓

One model + AdamW

Peak LR: 0.0022
One shared gradient and state

PACKED DECENTRALIZED

4 × 8,192-token local batches

Model 0	**Model 1**	**Model 2**	**Model 3**
m_i, v_i	m_i, v_i	m_i, v_i	m_i, v_i

Mix parameters each step

Peak LR: 0.0048
Evaluate $\bar{\theta} = \frac{1}{4}\sum_i \theta_i$

Each experiment runs on one GH200; “packed” describes the implementation.

Baseline topology: one peer per step, with alternating incoming offsets 1 and 2.

Adaptive consensus weakens mixing as LR decays

Mix local parameters, then apply local AdamW:

$$b_{i,t} = \sum_j W_{ij,t}\theta_{j,t}, \quad y_{i,t} = \gamma_t b_{i,t} + (1-\gamma_t)\theta_{i,t}$$

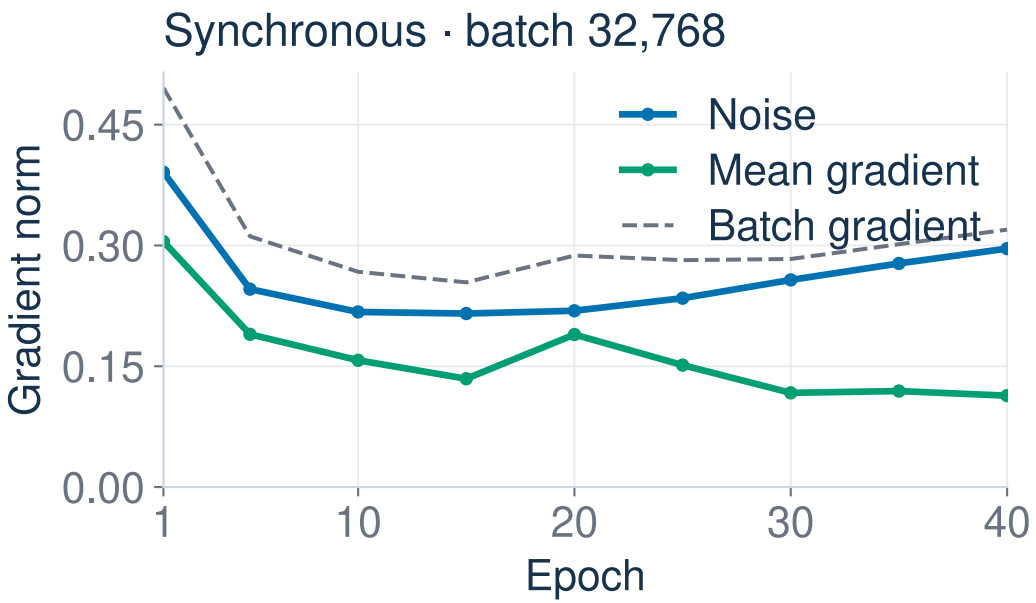
$$\gamma_t = \begin{cases} 1 & t < t_0 \\ \left(\frac{\eta_t}{\eta_{\max,\ \text{active}}}\right)^p & t \geq t_0 \end{cases}, \quad t_0 = \lceil 0.25T \rceil, \quad p = 2$$

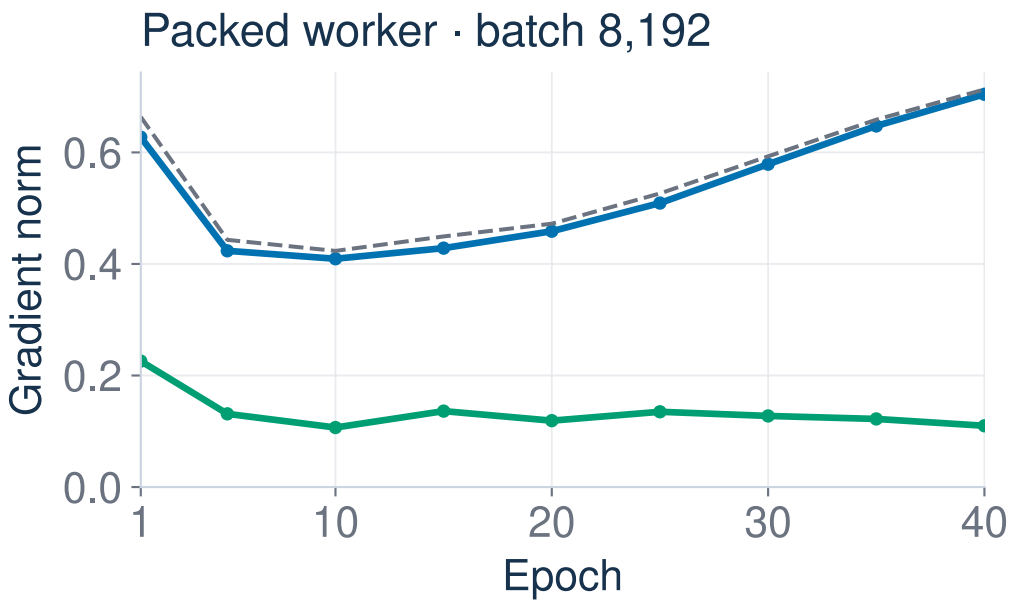
$\gamma = 1$ ordinary decentralized mixing $\gamma_{\text{end}} \approx 0.0122$ end of adaptive run

Local gradient → parameter mixing → local AdamW update

Gradients use pre-mixing parameters. Moments remain local. The LR normalizer is the maximum within the active interval.

Gradient noise grows relative to the mean gradient



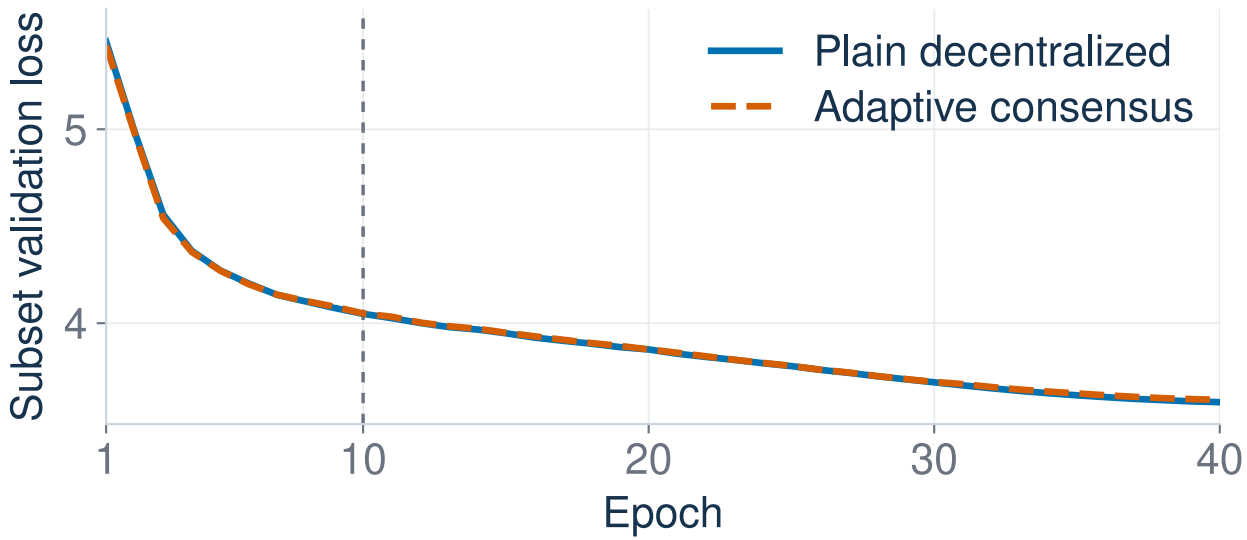


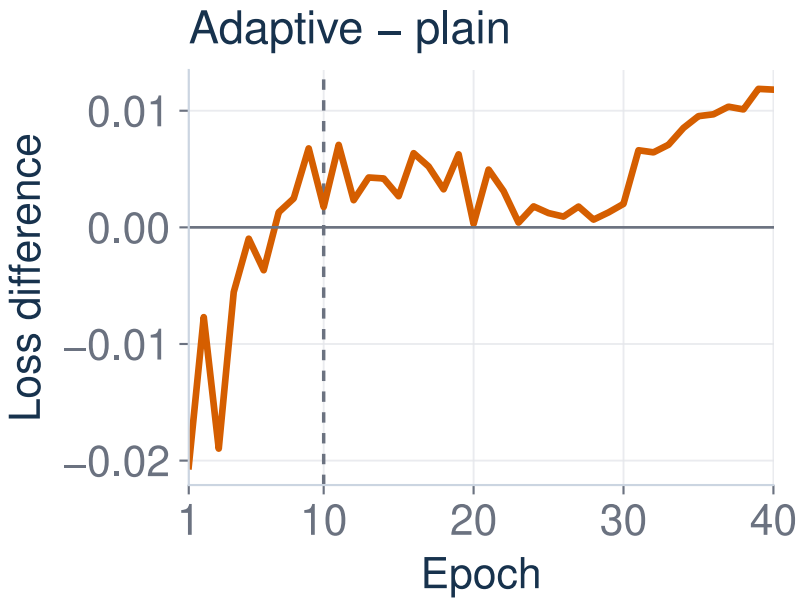
1.28 → 2.61 synchronous noise / mean

2.78 → 6.41 packed local noise / mean

Unseen training-cache data. Ratios compare average noise norm with mean-gradient norm.
Different local batch sizes and LRs: interpret the trend within each run.

Adaptive consensus shows no observed advantage

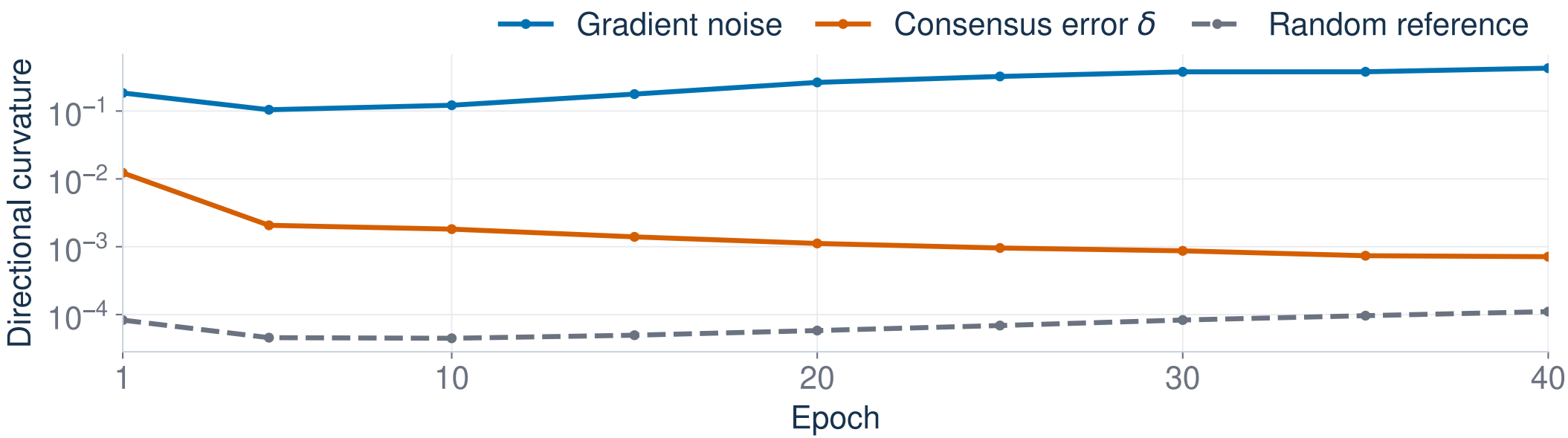




3.5959 plain **3.6062** adaptive **+0.0103** AC – plain

Callouts: final full-validation loss. Curves: fixed subset. Same topology, LR and budget; one seed.

Consensus error has lower curvature than noise

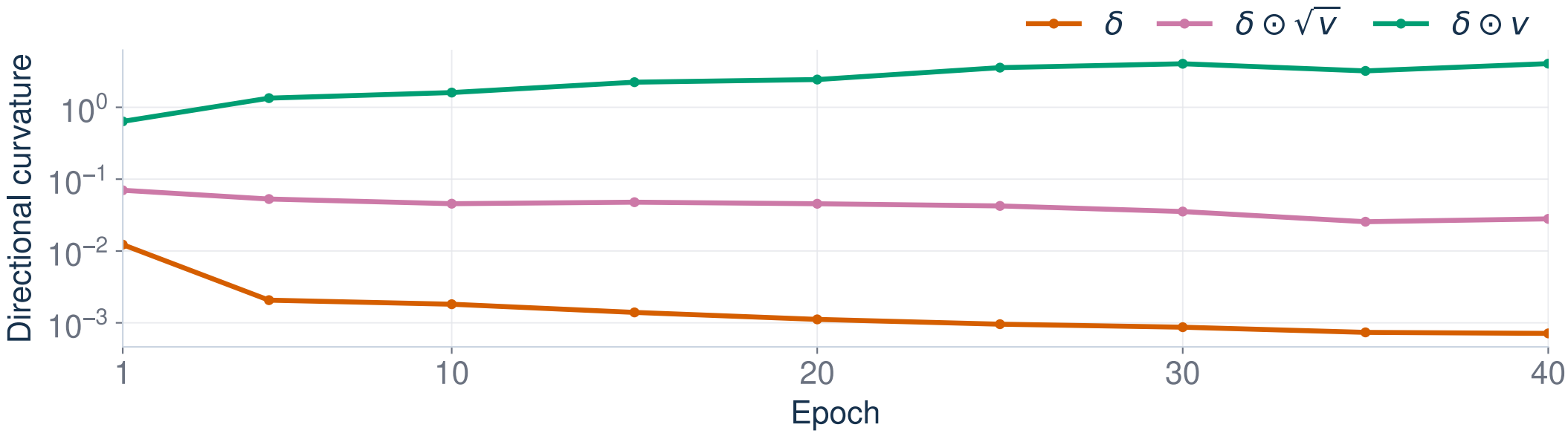


$$A(D;H) = \frac{\sum_i d_i^T H d_i}{\sum_i \|d_i\|^2}$$

236× below noise at epoch 20

$\delta_i = \theta_i - \bar{\theta}$. H is evaluated at $\bar{\theta}$ on unseen training-cache data.
Consensus error: local model minus global average. It scores above the random reference.

Weighting by v increases directional curvature

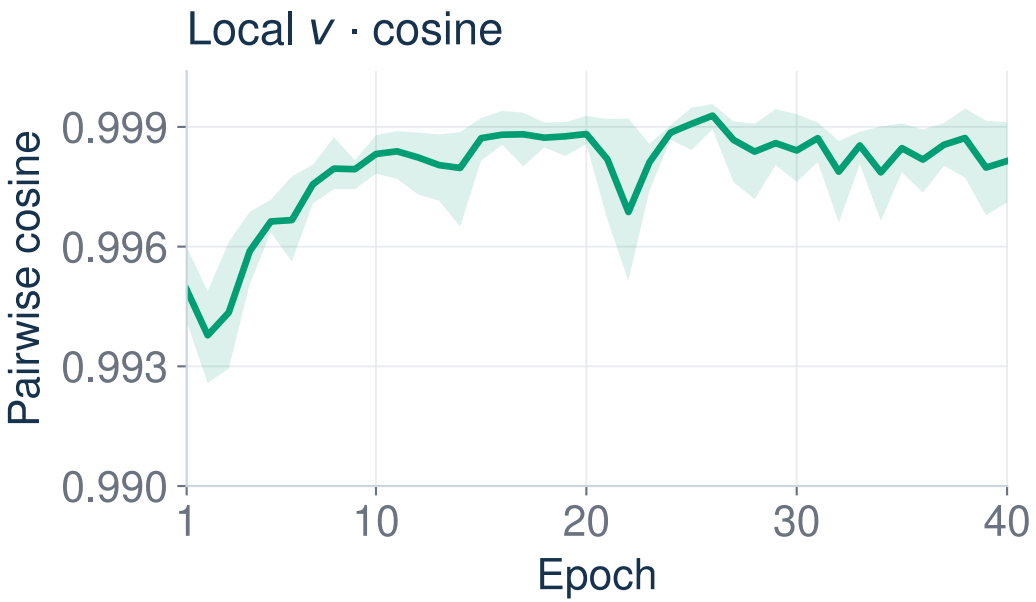


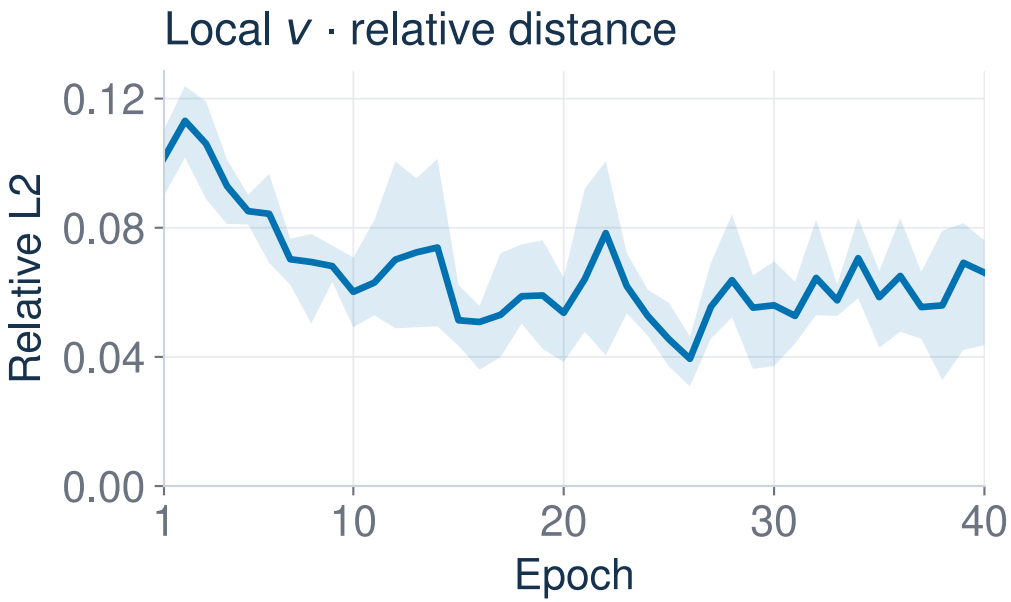
≈41× $\delta\sqrt{v}$ versus δ

≈2,178× δv versus δ

Epoch 20, unseen data; raw local second moments. These are transformed directions at saved AC checkpoints.

Local second moments are highly similar





0.9981
v cosine

0.0104
m cosine

0.0037
Adam-direction cosine

Epoch 40, all unique parameters. Bands show six-pair ranges. Global similarity allows coordinatewise differences.

Reshape the retained consensus error

TESTED DESIGN · COMPLETE TOPOLOGY · MAIN EXPERIMENT: α = 0.5

$$b_{i,t} = \overline{\theta}_t, \quad e_{i,t} = (1 - \gamma_t)\left(\theta_{i,t} - b_{i,t}\right)$$

$$s_{i,t} = e_{i,t} \odot v_{i,t}^{\alpha}$$

$$y_{i,t} = b_{i,t} + \left(\frac{\|e_{i,t}\|_2}{\|s_{i,t}\|_2}\right) s_{i,t}$$

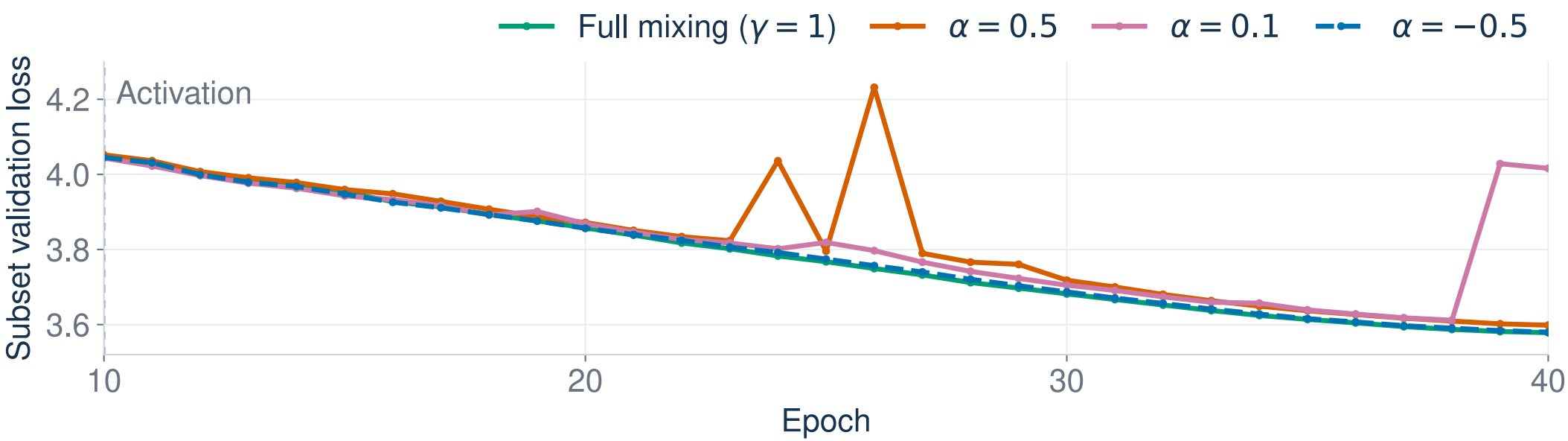
Then apply the same local AdamW update.

$v_{i,t}$: previous-step raw local moment

Normalize across all parameters per worker

$\odot$ denotes elementwise multiplication. Reweighting preserves $\|e_i\|_2$; the global average can shift. Zero transformed direction falls back to the original consensus error.

Moment-scaled training does not beat full mixing



3.5799
Full mixing

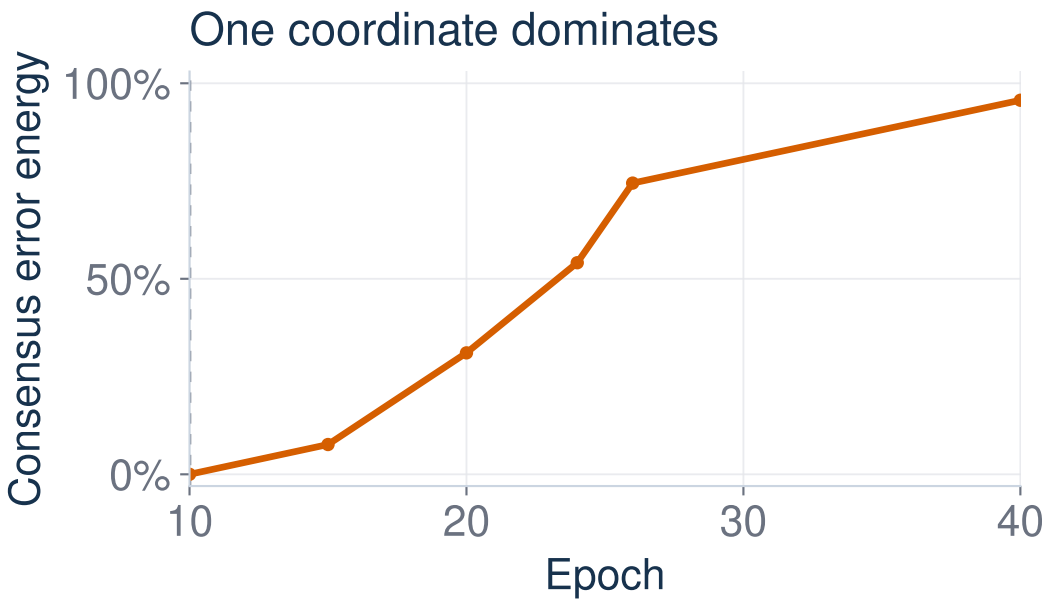
3.5827
$\alpha = -0.5$

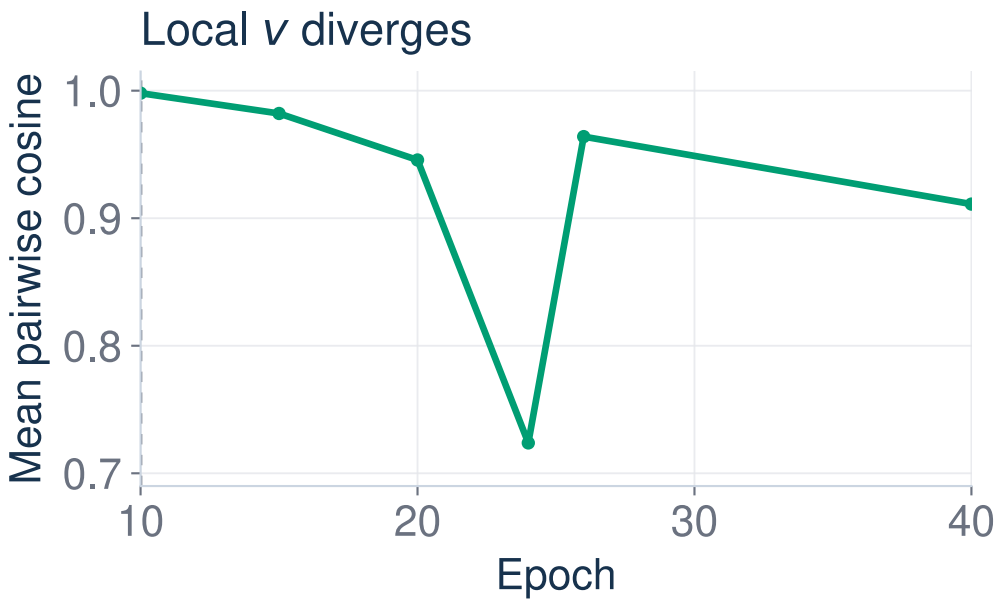
3.6002
$\alpha = 0.5$

4.0264
$\alpha = 0.1$

Final full-validation loss; one seed. Curves use the subset.
This review lacks the complete-topology, all-parameter adaptive $\alpha = 0$ control.

A fixed norm can hide extreme concentration





$$\frac{e_j^{(k)}}{e_l^{(k)}} = \frac{e_j^{(0)}}{e_l^{(0)}} \left(\frac{v_j}{v_l} \right)^{k\alpha}$$

95.7% in one coordinate

Illustration: fixed positive v, two symmetric workers, complete topology, no optimizer forcing.
Callout: worst worker, epoch 40. Repeated reweighting amplifies coordinate ratios.

A better diagnostic does not guarantee better training

1. Noise becomes more prominent relative to the mean gradient.
2. Moment weighting improves curvature on saved AC states.
3. Repeated weighting of consensus error concentrates its energy.

PROPOSED AT THE REVIEW STAGE · NO TRAINING EVIDENCE IN THIS REVIEW

Shape fresh optimizer-update disagreement

Shared, bounded gains · common normalization

Preserve the mean update; let ordinary AC contract existing error.

Proposed evaluation: add the matched control and judge replicated validation results.
Scope ends at the moment-scaling review; subsequent trials are outside this presentation.